Jan Rybicki, David Hoover and Maciej Eder

Computational Stylistics and Text Analysis

1. Introduction

Computational stylistics and text analysis have a long, rich history. In retrospect, because of the nature of texts and the capabilities of computers, it seems quite predictable that they would be among the first applications of computers to the humanities. Many religious, literary, and historical texts are highly valued cultural achievements, and some of them have been analyzed for hundreds or even thousands of years. They also contain large numbers of highly significant and meaningful words and other textual features. Thus it seems natural that scholars should have moved quickly to enhance and augment their own description, characterization, and analysis of these rich cultural documents by harnessing the power of computers to store, search, count, and compare textual features. The rapid growth of the power of computers and the rapid increase in the availability of electronic versions of texts have revolutionized the scope and the kinds of analysis that can be performed. At their core, however, computational stylistics and text analysis have remained true to their origins, and continue to use the power of the computer to improve our understanding of texts.

"Computational Stylistics" seems a relatively transparent phrase, but it may be useful to pick it apart a bit. First, "computational" obviously and correctly implies the use of computers, but it leaves unexpressed the rather wide range of ways they can be used. Simple text searches, concordances, and textual manipulation and selection could be counted as computational, but

most practitioners would reserve the term in this context for some kind of statistical analysis, ranging from t-tests, to Principal Components analysis and cluster analysis, to Delta analysis, data mining of various kinds, support vector machines, topic modeling, and even neural networks. Most of these analytic methods have their origins in the closely related field of authorship attribution, and, given that many of them focus on stylistic differences, the two fields are sometimes difficult to distinguish (Craig 1999). The process of distinguishing authors emphasizes the importance of differences and similarities, and almost all of the analytic methods applied to texts are focused on detecting these differences and similarities.

1.1 Style

Style, the subject matter of "Stylistics" is, most broadly, simply a way of doing something. In the simplest case, it is an author's way of writing. In practice, however, the focus is on the effects of an author's style on his or her texts. In one recent formulation, "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (Herrmann, van Dalen-Oskam, and Schöch 2015: 16). It is widely assumed, though unprovable, that the features that constitute each author's style form a unique stylistic fingerprint, so that, if the correct features are chosen, any two authors can be distinguished from each other. Considered as a property of texts, style can also be extended to apply to what can loosely be called genres (the Gothic novel, epic poetry, satire, narrative, drama), literary-historical periods (Victorian, Romantic), chronological divisions within an author's career (early and late Henry James), or to variations within a single text (the "voices" of different characters or narrators, for example), among many other possibilities.

Style is chiefly linguistic, though in some cases graphological features, the layout or arrangement of text, and even the physical characteristics of a text may contribute to a style. In addition to the obviously linguistic elements of style, such as vocabulary, grammar, morphology, phonology, and figures of speech, most practitioners would include broader characteristics, such as world view, theme, and tone, as potential elements for analysis (for an excellent checklist of stylistic features, see Leech and Short 2007: 61ff). Style is also patterned and distributed. Local and unique stylistic features can be important, but a recognizable style normally involves some kind of repetition, consistency, or pattern.

1.2 Stylistics

Stylistics is essentially comparative, even if the comparison is not always explicit. Almost all statements about a style imply a comparison; for example, even the seemingly simple statement that Faulkner's style is marked by long sentences implies a contrast with the lengths of the sentences of other authors. Long compared to what? Although the question of what norm is appropriate for the comparison remains vexed, the widespread availability of electronic texts and corpora of texts has made defensible choices and the creation of specialized corpora to use as norms easier to make. Pattern, distribution, and comparison obviously invite a computational approach. Indeed, computation analysis is the only practical way to analyze extremely frequent textual characteristics, or to study unreadably large collections of texts.

Despite the variety of stylistic features, it is fair to say that the overwhelming majority of computational stylistic analyses have involved words, though word n-grams (sequences of words) have recently become increasingly popular features to analyze. Not only are words (and

n-grams) easily identifiable and countable, compared to figures of speech, themes, or syntactic patterns, they are also much more frequent than most other textual characteristics and are obviously, though not unproblematically, meaningful (unlike, for example, sequences of letters or parts of speech). A "word" seems an intuitively simpler concept than it is in practice, and various decisions about how to identify and count the words of a text are defensible under different circumstances. For the purposes of computational stylistics, a word (type) is normally defined as any unique sequence of alphanumeric characters that is not interrupted by a space, or by any punctuation mark except the apostrophe or hyphen. (A *type* is a unique form, while a *token* is an individual occurrence of a type: the previous sentence contains two tokens of the type "or".) Unfortunately, this definition does not distinguish homographic forms like the noun and verb meanings of *desert,* but experience has shown that computational stylistics is robust enough that the resulting errors in counting do not seriously distort analysis.

1.3 Text Analysis

Text Analysis is a close relative of computational stylistics, but with a wider range and a heavier emphasis on analysis. While computational stylistics has focused almost exclusively on literary texts, text analysis has been applied to many other kinds of texts, from political speeches to blogs, from historical documents to tweets, from legal documents to the sacred texts of religions, from letters to philosophical treatises, from poetry to programming. Perhaps the most obvious further difference between computational stylistics and text analysis is that the latter is more likely to focus on meaning and content. Nevertheless, almost all of the methods of text analysis

have also been applied to questions of literature, authorship, and style.

It would be an exercise of folly to attempt an introduction here to data mining, topic modeling, sentiment analysis, semantic analysis, neural networks, part-of-speech analysis, word-frequency analysis, and of the wide range of statistical analysis techniques that have been applied to the dozens of different textual features that have been analyzed. Instead, it seems more useful to approach computational stylistics and text analysis by focusing on problems at three different scales: microanalysis, middle-distance analysis, and macroanalysis or distant reading.

The first analysis on the micro scale is an authorship problem involving a collaboratively written text, *The World's Desire* by H. Rider Haggard and Andrew Lang, using a popular recent technique called Rolling Delta. This is followed by an analysis of the voices of the six narrators in Virginia Woolf's *The Waves*. Middle-distance analysis is represented by a modification of John Burrows's Zeta (Burrows 2006) that examines the vocabulary of more than 350 high-stakes exit essays written by American high school students. Macroanalysis is demonstrated by turning to the chronological signal visible in much larger corpora: in this case, in 1000 English novels from Swift (Jonathan) to James (E. L.).

2. Microanalysis, or Zooming into a Single Text

Empirical investigations in the field of computational stylistics and text analysis are, as noted above, almost exclusively focused on comparison: to reliably describe a given text's statistical characteristics, in a vast majority of cases one compares the text to other texts collected in a comparison corpus. From this perspective, a single text, perceived in a context of similar or not-

so-similar texts, becomes a monadic entity *per se*. Even if such a text is further divided into smaller samples (see e.g. Kestemont, Moens, and Deploige 2015), the main goal of finding relations between discrete textual entities (works) continues to be the main focus.

This approach assumes that a (literary) work is a monolith, which is not always true: an epistolary novel might consist of multiple stylistic registers, a Menippean satire might combine sections of epic poetry, tragedy, and philosophical prose, a collaboratively written work might contain two or more independent authorial voices, and so forth. In such cases, capturing an average stylistic profile from the text in its entirety is certainly not the optimal scenario. Arguably, much more can be observed when such a text is divided into segments and treated independently. One of the possible applications of this approach is discussed below, where Virginia Woolf's *The Waves* is dissected according to particular characters' voices; another application involves chunking the input text into consecutive samples, or equal-size blocks of *n* words (tokens), that are then measured as independent, yet sequentially ordered, samples.

2.1 *The World's Desire*

Pioneering work in sequential stylometry was presented in a study on the authorship of *Walewein* (van Dalen-Oskam and van Zundert 2007), in a comparison of three disputed English prose texts (Burrows 2010), and in a study of *The Tutor's Story*, written collaboratively by Kingsley and Malet (Hoover 2012). The sequential methodology evolved into the Rolling Delta method (Rybicki, Hoover, and Kestemont 2014), later extended and generalized as Rolling Classify (Eder 2015a). This method will be used here to assess the nature of collaboration between Henry Rider Haggard and Andrew Lang on *The World's Desire*, first published in 1890.

2.1.1 Background

Henry Rider Haggard (1856-1925) is the author of several adventure novels, among which the bestsellers *She* (1887) and *King Solomon's Mines* (1885) attracted a good deal of attention. Andrew Lang (1844-1912), a poet, novelist, literary critic, and folklore scholar, earned his fame as a translator of Homeric poems. *The World's Desire*, a classic fantasy novel written collaboratively by the duo, not particularly long (*ca*. 85,000 words), is a story of the hero Odysseus, who returns home to Ithaca after his journey: instead of finding his home at peace, however, he is involved in several new adventures. The plot of the novel as well as its mythological background was set by Lang, while Haggard contributed his imagination and style. From the correspondence of the two writers, we know that Haggard had written a first draft, entitled *The Song of the Bow*, that was later reworked by Lang. Haggard then took over and wrote a great share of the text. In Haggard's own words:

> Roughly the history of this tale […] is that Lang and I discussed it. Then I wrote a part of it, which part he altered or rewrote. Next in his casual manner he lost the whole MS. for a year or so; then it was unexpectedly found, and encouraged thereby I went on and wrote the rest. […] The MS. contains fifty-three sheets at the beginning written or re-written by Lang, and about 130 sheets in my writing, together with various addenda. (Haggard 1926)

It is assumed that Haggard actually wrote most of the novel except the first four chapters, which were written entirely by Lang. Working on the first drafts of the novels, the two authors were

aware of stylistic differences between them. Haggard quotes Lang's letter, which confirms his habit of depreciating his own work:

> Nov. 27th. The typewritten "Song of the Bow" has come. The Prologue I wrote is better out. It is very odd to see how your part (though not your *chef d'oeuvre*) is readable, and how mine—isn't. (Haggard 1926)

2.1.2 The Two Authors in *The World's Desire*

The work by Haggard and Lang seems to be a perfect case study of mixed authorship of a single text. To tell its authorial voices apart, the Rolling Classify was applied. First, the goal was to compile a reference corpus containing authorial profiles of both Haggard and Lang. Out of an extensive list of Haggard's works, 10 novels and 2 collections of short stories were selected to train a Haggardian profile: *Cetywayo and His White Neighbours* (1882), *Allan's Wife and Other Tales* (1887), *Allan Quatermain* (1888), *Colonel Quaritch, V.C.* (1888), *Cleopatra* (1889), *Beatrice* (1893), *Black Heart and White Heart* (1900), *Ayesha: The Return of She* (1905), *Benita* (1906), *The Yellow God* (1908), *Child of Storm* (1913), *Allan and the Holy Flower* (1915). When it comes to Lang, a similar selection of 12 novels and short stories collections was compiled: *Much Darker Days* (1884), *In the Wrong Paradise and Other Stories* (1886), *He* (1887), *The Gold of Fairnilee* (1888), *Prince Prigio* (1889), *The Green Fairy Book* (1892), *Prince Ricardo of Pantouflia* (1893), *The Disentanglers* (1902), *The Crimson Fairy Book* (1903), *The Olive Fairy Book* (1906), *The Brown Fairy Book* (1904), *The Lilac Fairy Book* (1910).

The above representative text samples are used to train a model using one of the supervised classification techniques, namely, Support Vector Machines. The testing procedure

starts with chunking *The World's Desire* into consecutive samples, or equal-size blocks of 5,000 words, with an overlap of 4,500 words, to achieve a dense sampling rate. Next, the support vector machine classifier is applied sequentially to the particular samples, which are checked against the training set, in order to identify the most similar authorial profile. The final stage of the analysis involves a graphical representation of stylistic changes throughout the chunked text. To this end, horizontal stripes are used, which are colored according to the assigned class.

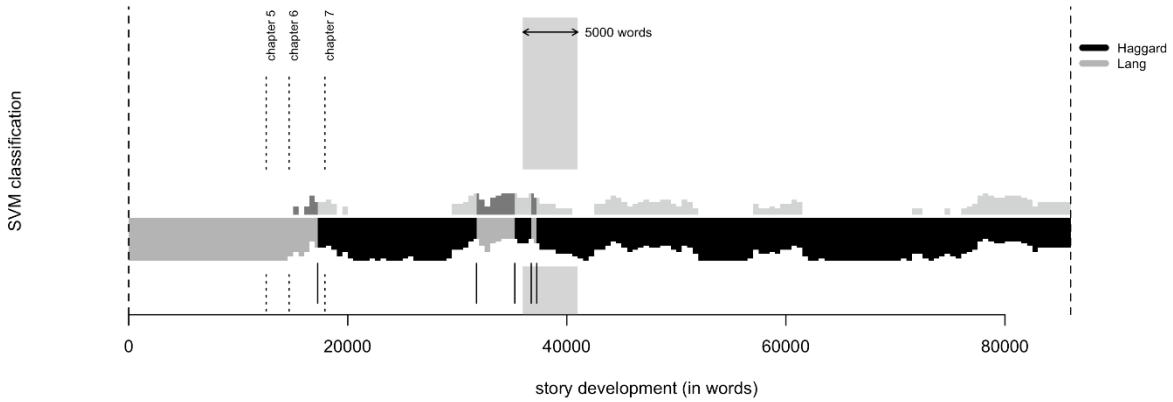

Figure 1. Sequential analysis of The World's Desire by Haggard and Lang: Rolling SVM and 100 MFWs

In Fig. 1, the results of the Rolling Classify technique applied to *The World's Desire*, using 100 MFWs, are shown. One can easily observe a stylistic takeover in the first part of the text. The break point takes place in the middle of the sixth chapter. Also, some sections by Lang seem to appear in the central part of the novel. However, these evaporate when a different MFW stratum is tested.

In Fig. 2, one can observe the behavior of *The World's Desire* when 500 MFWs are analyzed. This time, Haggard's signal shows up for a short moment in the first chapters of the novel. The picture is once more slightly different when 1000 MFWs are taken into consideration (Fig. 3). It is quite clear that for very long vectors of frequent words, the distinction between two authorial voices in the sixth chapter is the only takeover that can be observed.

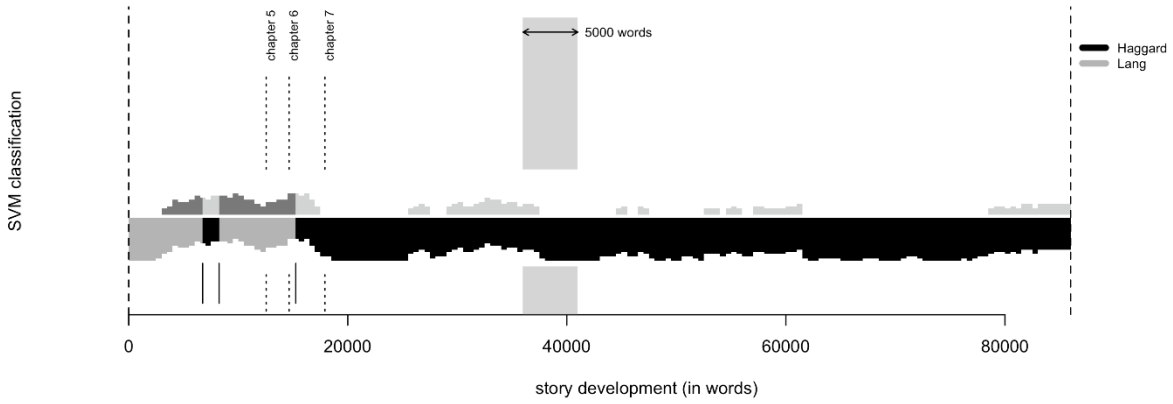

Figure 2. Sequential analysis of The World's Desire by Haggard and Lang: Rolling SVM and 500 MFWs

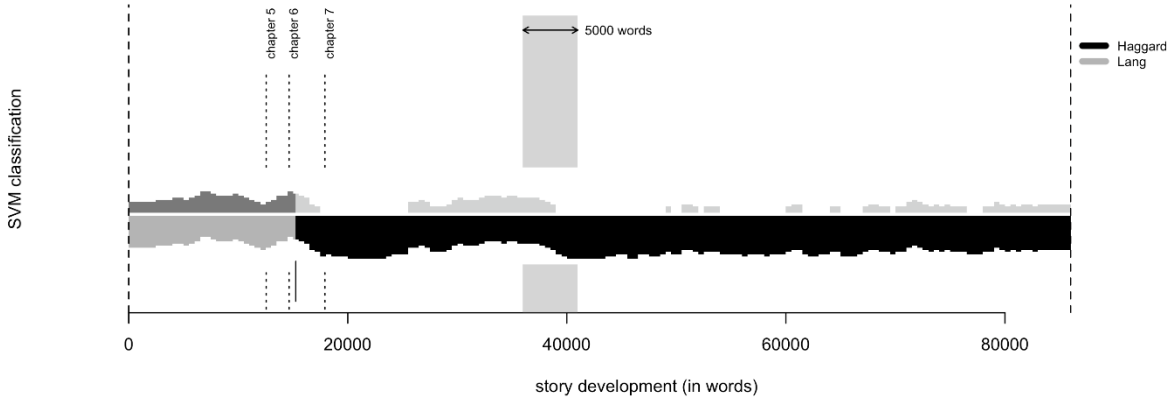

Figure 3. Sequential analysis of The World's Desire by Haggard and Lang: Rolling SVM and 1000 MFWs

Comparison of Figures 1-3 leads to the conclusion that the mixed authorship has a form of a sudden takeover rather than a mixture of interwoven authorial voices. However, it is much more difficult to explain the clutter that appears in different segments of the novel depending on the input parameters of the model. This observation is confirmed by a series of similar tests using different classifiers and different style-markers, such as the most frequent word 2-grams (word pairs). At this point, one of the most difficult problems of text classification arises, namely, the distinction between an actual signal and wrong decisions of the classifier (also referred to as false positives). Since this problem goes far beyond the scope of this chapter, it will not be discussed in detail. However, an intuitive and relatively simple way of filtering out the false positives is to perform a series of similar yet not identical tests, in which the parameters of the model (e.g., the number of MFWs) are modified. Patterns that appear despite differences in MFWs tested seem to suggest the existence of a signal, while any ephemeral clutter in the results might simply mean false positives. If this rule-of-thumb is true, *The World's Desire* has one takeover only, which falls in the middle of the sixth chapter.

2.2 Virginia Woolf's Character Voices

The Waves, Virginia Woolf's most experimental novel, consists of alternating soliloquies or monologues by three male and three female characters, from childhood through middle age, each clearly indicated by a simple speech-reporting phrase like "said Bernard" or "said Jinny."

Woolf's technique has invited considerable critical comment about what axes of difference or unity characterize the novel, as Stephen Ramsay has noted:

> Are Woolf's individuated characters to be understood as six sides of an individual consciousness (six modalities of an idealized Modernist self?), or are we meant to read against the fiction of unity that Woolf has created by having each of these modalities assume the same stylistic voice? (2011: 10).

Ramsay's claim that it would be a mistake to treat the question of whether the six voices are the same or different as one that can be answered has been disputed (Hoover, forthcoming; see also Hoover 2014 and Plasek and Hoover 2014). However, it seems worthwhile to leave the polemic aside, here, and look a bit more closely at the voices. (For a thorough recent discussion of the various views about the similarities and differences among the voices in *The Waves,* see Balossi, 2014: Chapters 1-2.)

2.2.1 Distinguishing the Six Voices

Testing the similarities and differences among the six voices is not as simple as it might seem. Because they are so obviously different from the monologues, it seems prudent first to eliminate the sections of third-person narration that begin each chapter and to remove all quotations from other characters from the monologues, so as to analyze only each character's voice (as does Burrows 1987: 191, 205-07). More problematically, the lengths of the six monologues vary from Susan's 6,067 words to Bernard's 32,664. The final chapter of the novel, which is all in Bernard's voice, begins "Now to sum up," showing that it is likely to be quite different from the rest of the novel. This chapter has been excluded from the analysis, as it was in three previous

analyses of the novel (Burrows 1987: 206; Ramsay 2011; Balossi 2014: 84). Unfortunately, this still leaves the numbers of words by each character quite unbalanced: Bernard, 16460; Jinny, 6281; Louis, 8694; Neville, 9958; Rhoda, 8401; Susan, 6067. To give each character the same weight, each monologue has been reduced to the length of Susan's. Simply taking the first 6,067 words of each, however, would mean that only Susan's whole life would be represented, so the lines of each monologue were randomized and each was cut to 6,067 words. After creating a word frequency list based on the six equal parts, a series of cluster analyses was performed, based on the 100, 200, . . . 1,000 most frequent words of this list on the six equal parts and all the remaining text in sections of about 3000 words. All the analyses correctly group all of the sections by a single character except the one based on the 500 most frequent words; a representative analysis based on the 900 most frequent words is shown in the cluster analysis in Fig. 4[1]. A similar analysis, along with others based on word 2-grams (sequences of two words) and words selected on the basis of consistent occurrence rather

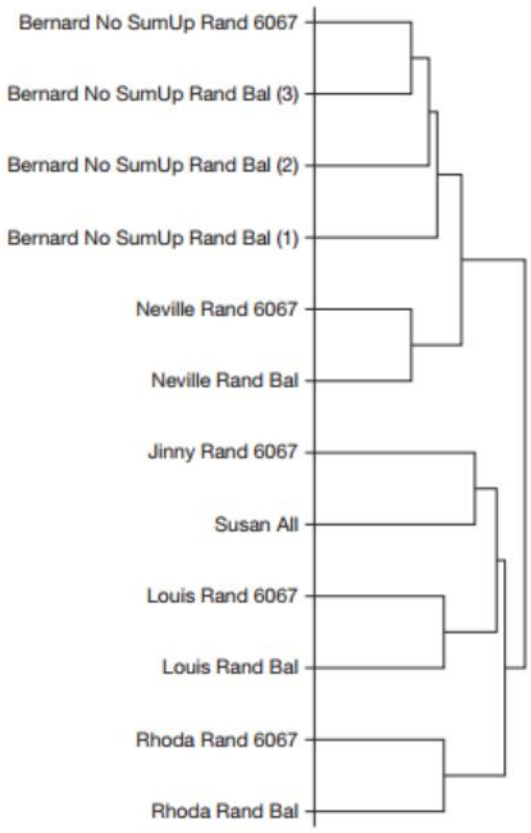

Figure 4. Randomized 6,067-Word Sections of The Waves – 900MFW

[1] Cluster analysis is an exploratory method of analysis that is often used in authorship studies. It compares the similarities and differences among the frequencies of all the words being analyzed in all the texts. The texts that use the words in the most similar way are grouped together in such a way that the more similar two texts are, the closer to the left of the graph they join together. In Fig. 4, the two most similar sections are the top two by Bernard (the vertical proximity of sections like Neville Rand Bal and Jinny Rand 6067 is not meaningful).

than high frequency also clearly distinguish the six voices in *The Waves* (see Hoover, forthcoming, for details). In spite of the fact that these six voices are all obviously versions of Woolf's voice, they are much easier to distinguish than are sections of texts by some pairs of authors.

2.2.2 Age and Gender and the Six Voices

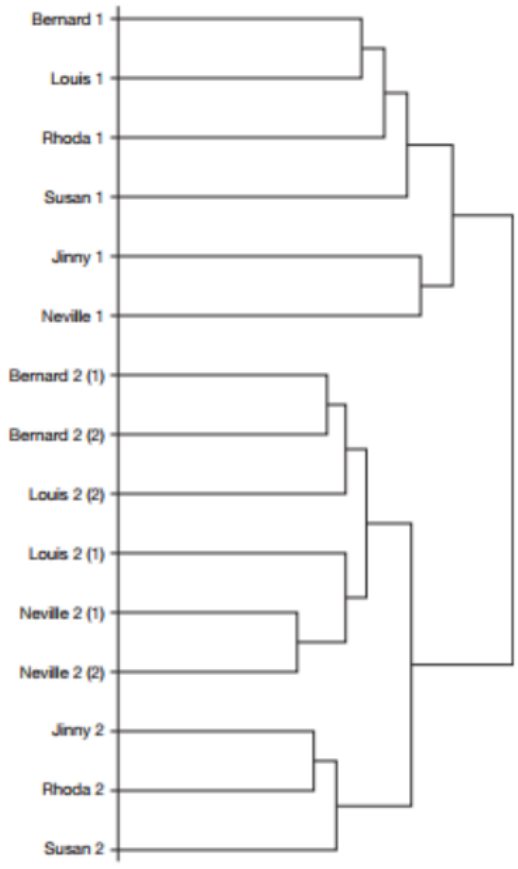

Figure 5. Chapters One and Two of The Waves, 500MFW

Given how distinctive the six voices are when analyzed with the lines randomly organized, it may be surprising that, at the same time, she also manages to distinguish the young voices from the older ones. In chapter one, the six are young children and in chapter two they go off to boarding schools. By chapter three they are entering college, and at the end of the novel the five surviving characters are middle-aged. A cluster analysis based on the 500MFW of the first and second chapters is shown in Fig. 5. Note that only Bernard's part, at 1,598 words is longer than 1,000 words, Jinny's is only 405 words, and Neville's only 505; most analysts would consider these too short for reliable analysis. Nonetheless, all six of the first chapters cluster separately from those from the second (sections from the second chapter that are longer than 2,000 words have been divided into one section of 1,000 words and a second section consisting of the remainder). The two sections by Bernard and

Neville (though not Louis) also cluster together, and the sections of chapter two by the three female characters cluster separately from those of the male characters, suggesting that, at least when they are children, Woolf has also created a gender split in their language. Cutting these chapters into still smaller sections (405-799 words) shows just how carefully Woolf has distinguished the voices in her novel. As Fig. 6, based on the 700MFW, shows, only the second half of Rhoda's chapter one fails to cluster with the rest of the chapter one sections at the top of the graph, and the sections of Rhoda, Jinny, and Susan from chapter two cluster together, though those of Bernard, Neville, and Louis cluster only partially. What is extraordinary here is how well these very short sections group together both by age and by character, and, except for two of the sections by Louis, also by gender. (For a more detailed analysis of the younger and older voices, based on very different methods, see Balossi 2014: chapter 6 and Appendix E; for a discussion of the strengths of authorship, genre, gender, and other signals in texts, see Jockers 2013: chapter 6).

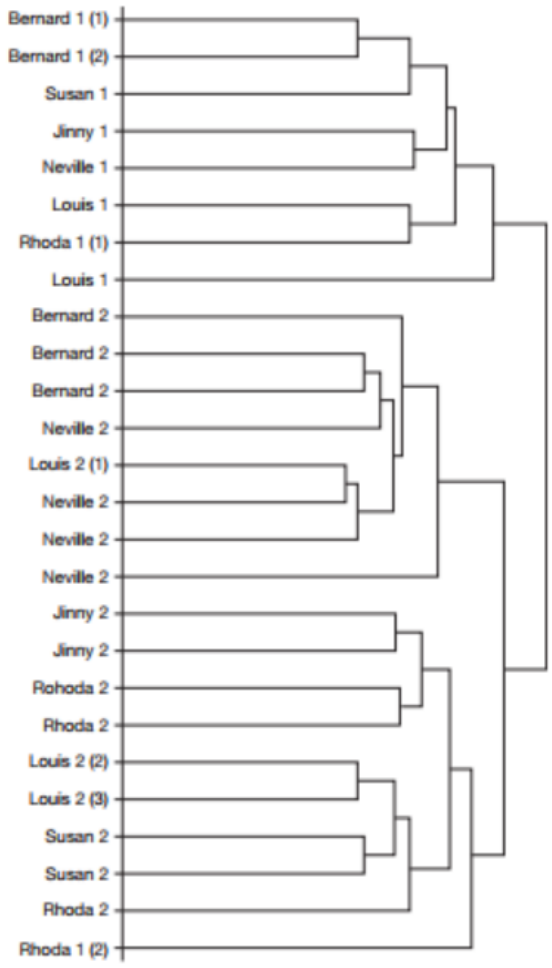

Figure 6. Chapters One and Two of The Waves, 700MFW (sections of 405-799 Words)

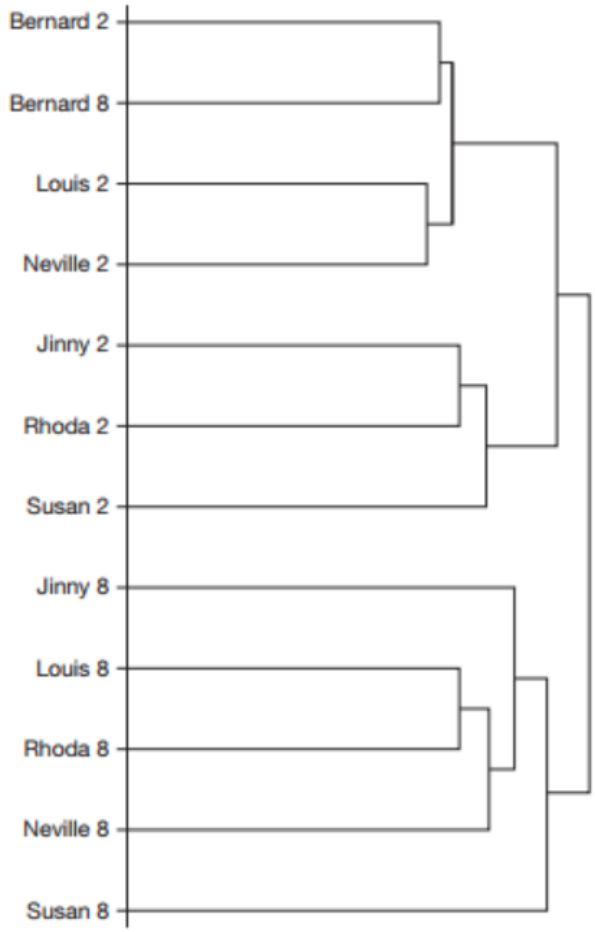

Figure 7. Chapters One and Two of The Waves, 700MFW (sections of 405-799 Words)

One more analysis will show that Woolf also distinguishes the voices of her characters as boarding school students in chapter two from their voices as adults in chapter eight. As Fig. 7 shows, with the exception of Bernard's monologue from chapter eight, all of the monologues from chapter two cluster separately from those of chapter eight. An examination of the monologues from chapter two suggests that, even at a young age, Bernard's style is more mature and complex that those of the other characters. Although the gender separation of chapter two naturally reappears here, it disappears entirely in the monologues from chapter eight (looking back at Fig. 4 shows that the randomized parts also fail to group by gender). Obviously, there is no space here for any approach to a full investigation, but these results show convincingly that the methods of computational stylistics can be valuable for exploratory studies, not least by suggesting productive new possibilities for further analysis and discussion.

3. Middle-Distance Analysis: High Stakes Writing Exams

Consider now a middle-sized group of very different texts: high-stakes exit-level writing exams. Although these texts lack any significant intrinsic value of their own, the tools of computational stylistics and text analysis can still produce revealing and worthwhile results. The texts to be analyzed here are a set of 366 essays written by North American high school students in the context of state-wide exit-level writing exams, administered in the final year of high school. These essays were collected and analyzed in a study of the idea of voice in writing assessment (Jeffrey 2010). The essays are not ideal for computational text analysis because they are quite short, averaging only about 430 words and ranging from 128 to 1307 words (only 21 are shorter than 200 or longer than 800 words). They also come from 39 states, and are responses to many different prompts that call for writing in a variety of genres; for example, analytic, narrative, argumentative, explanatory, and informative. If analyzing texts can achieve interesting results under these unfavorable conditions, we can expect excellent performance on longer texts under more favorable conditions.

3.1 Methodology

The method demonstrated here is a variant of Burrows's Zeta (Burrows 2006; Hoover 2007), as modified by Craig and Kinney (2010) and further modified in Wide-Spectrum analysis (Hoover 2013). The method is especially useful for characterizing the vocabularies of any pair of authors, genres, texts, or indeed any pair of text collections that can be divided unambiguously into two classes. Here, the first comparison will be between low-scoring and high-scoring passing exams (failing exams tend to be very short and defective). Although there are undoubtedly differences of many kinds between the two groups, here we concentrate on what

kinds of consistent vocabulary differences, if any, exist between the low-scoring and high-scoring essays.

Wide-spectrum analysis, unlike most methods of text analysis, is based on consistency of use rather than frequency, and its calculation is simple and straightforward. For example, assume there are 200 texts, approximately the same length, by two authors, 100 by each author. Assume further that the word "eyes" is present in 62 (.62) of the texts by the first author and absent from 92 (.92) of the texts by the second author. These percentages are added together to yield a distinctiveness score for "eyes" of .62 + .92 = 1.54. Although distinctiveness scores can range from two (100% presence plus 100% avoidance) to zero (0% presence plus 0% absence), in practice scores above 1.5 are strongly characteristic of the first author and those with distinctiveness scores below .5 strongly characteristic of the second author. Note that quite different distributions can produce similar distinctiveness scores. For example, a word that is present in 77% of the texts by one author and absent from 77% of texts by the other would also have a distinctiveness score of 1.54.

Because of the wide range of sizes in the essays to be analyzed and the variety in prompts and genres, all the low-scoring essays have been combined into one text and all the high-scoring essays into another text and then the lines of each combined text have been sorted in random order. The final 14,000 words of each randomized text have been reserved for testing, leaving the rest of the texts for training purposes. Finally, to avoid basing the analysis on topical words, or words that are frequent because of specific prompts used in states with large numbers of essays, the word list has been manually culled by removing proper names (names of states and character names from text-based prompts, for example) and other topical words. The randomized training

and testing texts were then cut into blocks of 2000 words and analyzed in the Wide-Spectrum spreadsheet, which automates the process of comparing the two main sets of texts, calculates a distinctiveness score for all the words, and sorts those above a neutral score of one from high to low and those below one from low to high. The sheet also collects the most distinctive words for each group in order of distinctiveness, graphs the results, and prepares them for further graphing.

3.2 Vocabulary and the Evaluation of High School Writing

The results of the analysis described above are presented in Fig. 8.

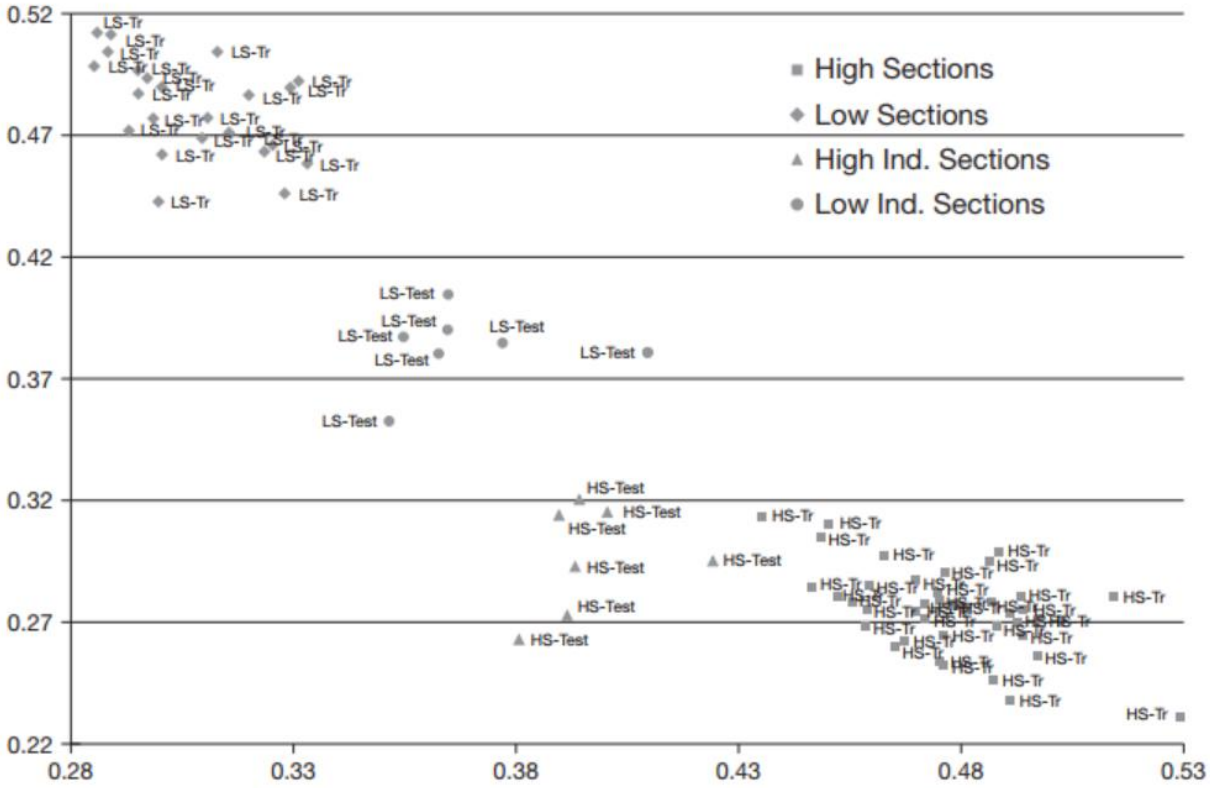

Figure 8. High and Low Scoring Essays

The horizontal axis in Fig. 8 indicates the percentage of the word types in each section that are characteristic of the training sections of the high-scoring exams and the vertical axis shows the percentage of the word types in each section that are characteristic of the training sections of the low-scoring exams. (A word type is defined here as a unique spelling; each 2,000-word section typically contains only about 700-800 types because many common words are repeated.) For example, for the high-scoring training section at the bottom right of the graph, only about 23% of the 820 different types are characteristic of low-scoring exams, while almost 53% are characteristic of high-scoring exams. It is easy to see that this method does an excellent job of categorizing the sections that were held out for testing, in spite of the fact that the test sections had no part in creating the word lists on which Fig. 8 is based. All fourteen of the high-scoring and low-scoring sections are much closer to the appropriate training texts than to the opposite ones. (Tested sections do not usually fall within the clusters of training texts because they contain many words that are not in the training sets, so that the percentages of types that are characteristic of each group are lower than for the training texts.)

Although Wide-spectrum analysis does an excellent job of categorizing the test texts here, its real value is in characterizing the texts themselves. Below are the fifty most distinctive words for the two groups, unlabeled so that readers can try their hands at identifying which list is characteristic of the high-scoring essays and which is characteristic of the low-scoring ones.

big, lot, might, makes, stop, maybe, talking, major, example, picture, conclusion, older, sometimes, I'm, try, bad, tell, trouble, hurt, harder, anywhere, got, end, give, remember, problems, affect, car, understand, times, goes, cars, younger, basketball, nice, anything,

wouldn't, saying, public, everybody, kids, decision, weather, doing, mom, teen, extending, normally, technique, someone

completely, must, eyes, desire, physical, constantly, individual, far, stress, mind, precious, human, light, involved, simply, rather, understanding, fit, build, led, ever, reader, air, changes, beginning, red, responsibility, order, nothing, began, increase, eventually, future, views, clock, merely, which, college, true, difficult, actually, science, once, learning, one's, ability, walk, although, various, aware

Most readers correctly identify the first list as words characteristic of the low-scoring essays and the second list as words characteristic of the high-scoring essays. There is no space for a full analysis of the differences in the lists, but two important dimensions are formality and specificity (see Jeffrey 2010 and Jeffrey, Hoover, and Han 2013 for more details). The low-scoring essays use a much more casual and informal vocabulary, most noticeable in *big, lot, I'll, bad, got, wouldn't, kids,* and *mom,* while the high-scoring essays use more formal words like *desire, physical, individual, precious, responsibility, science, one's,* and *various.* The low-scoring essays also tend to use vague and unspecific vocabulary like *lot, sometimes, bad, anywhere, affect, nice, anything, everybody, doing, someone,* while the high-scoring essays tend to be more specific, using words like *completely, constantly, stress, human, build, increase, difficult,* and *learning.*

An examination of the 1,000 most distinctive words for each group confirms these trends. For example, the first list above contains two contractions and the second none, and there are just

three contractions among the 1,000 most distinctive high-scoring words, compared to twenty-two among the 1,000 most distinctive low-scoring words. Along with *anywhere, anything, sometimes,* and *someone* in the list above are *somebody, someday, someone's, something, sometime, whenever,* and *whatever* among the 1,000 most distinctive. The nouns in the two lists are also revealing:

Low-scoring nouns:

lot, example, picture, conclusion, trouble, end, problems, car, times, cars, basketball, kids, decision, weather, mom, teen, technique

High-scoring nouns:

eyes, desire, individual, stress, mind, light, understanding, reader, air, changes, beginning, responsibility, order, future, views, clock, college, science, ability

It *isn't hard* to *use* a *lot* of *nice* low-scoring *examples* to *give* a *picture* of *kids* who *at times* have *problems*, *athletes* who *really love cars, watching television,* and *playing sports like basketball,* and *want* to *talk* about the *weather* (words that are among the 1,000 most distinctive low-scoring words are in italics).

However, the high-scoring words *present* a *well-rounded individual*, a *reader* with a *passion* for *learning* and a *desire* for *future responsibility* who is *admired* by her *teachers, almost*

all of *whom acknowledge* her *leadership potential* and her *deep understanding* of *literature.* (words that are among the 1,000 most distinctive high-scoring words are in italics).

This kind of analysis provides a great deal of interesting evidence about what the exam graders value and do not value, though it would be dangerous and irresponsible to imagine that the two passages above fairly characterize the two groups of students. While there is no space here to pursue the analysis further, it should be clear that such evidence can be very profitably used in discussing and evaluating the process of high-stakes writing tests and their implications for education.

4. Massive Text Analysis of 1000 Novels

It seems that any quantitative analysis worth its salt should yield significant chronological differences in a 1000-novel corpus that extends from the times of Jonathan Swift to those of E. L. James, especially if it is based on such simple linguistic features as word frequencies. There is an obvious expectation of linguistic difference between the former's and the latter's images of, say, bondage in, respectively, *Gulliver's Travels into Several Remote Nations of the World* and *Fifty Shades of Grey*:

> I had the fortune to break the strings, and wrench out the pegs that fastened my left arm to the ground, for, by lifting it up to my face, I discovered the methods they had taken to bind me, and at the same time with a violent pull, which gave me excessive

> pain, I a little loosened the strings that tied down my hair on the left side, so that I was just able to turn my head about two inches.
>
> His deft fingers skim my back occasionally as they work down my hair, and each casual touch is like a sweet, electric shock against my skin. He fastens the end with a hair tie, then gently tugs the braid so that I'm forced to step back flush against him. He pulls again to the side so that I angle my head, giving him easier access to my neck.

Or is there? While no machine is needed to learn and discover the difference between the two texts, most of the differences between the two depictions – apart from what was then and what is now acceptable in print – seem to concern syntax rather than lexis: "I a little loosened the strings" sounds 18th-century, not 21st; but the narrator of *Shades*, had it been part of the game, would probably have said "I loosened the strings a little," and the words and their frequencies would have remained unaffected. And yet Fig. 9 shows the two texts at two different extremities in a network visualization of nearest-neighbor links between texts based on cluster-analyzed Delta distances between them based in turn on most-frequent word frequencies (Eder 2015b). While the texts by the two authors are bound by the strongest signal known to stylometry, that of authorship, the authorial groups for both groups clearly order according to general chronology.

More importantly, not only those two texts behave this way. Whatever is between and around their data points follows suit, and the entire network exhibits a very Morettian evolution of greyscale from black to light grey, from left to right and from early to late. There are departures from a purely linear sequence, of course, and some authors move and/or evolve

vertically rather than horizontally; but the overall phenomenon is unmistakable. But then this has already been noted a long time ago with other methods, different statistical tools, and other corpora (not to reach too far back, cf. Brainerd 1980, Burrows 1994, Opas 1996). It is perhaps more significant that, against the background of this great evolution, local evolutions in chronology can be observed in some authors – those, obviously, who are represented by more texts in this corpus than either J. Swift or E. L. James. The latter's more august last-name-sake, Henry, has already been shown to evolve in the context of his own work (Hoover 2007); but here, too, the chronological sequence of his works in Fig. 9 is almost perfect. Dickens evolves as well: while less linearly, his evolution seems to follow the general left-to-right trend. By contrast, Joseph Conrad's evolution seems to go against the grain: it is almost as if the Polish-born writer evolved his English towards a more "traditional" usage of most frequent words.

When the two evolutionary phenomena – that of an English literature more or less smoothly moving in a single direction with time, and of a similar movement within a single author – are put together, the picture becomes, perhaps paradoxically, less clear. When one compares the scale of the evolution of Henry James with that of the entire corpus, the former seems to be blown out of proportion. Obviously, there is no direct reflection of any quantitative distance between the texts, and the distances themselves are the combined results of the balanced forces of the gravitational pull of the networking algorithm; still, if the entire "stylistic drift" were to be blamed on historical-linguistic factors alone, that proportion would probably be better maintained. Perhaps most significantly, whatever it is that a network analysis of this kinds represents, it seems to accord quite well with combined literary and linguistic expectations. In simpler language, this network makes a lot of sense.

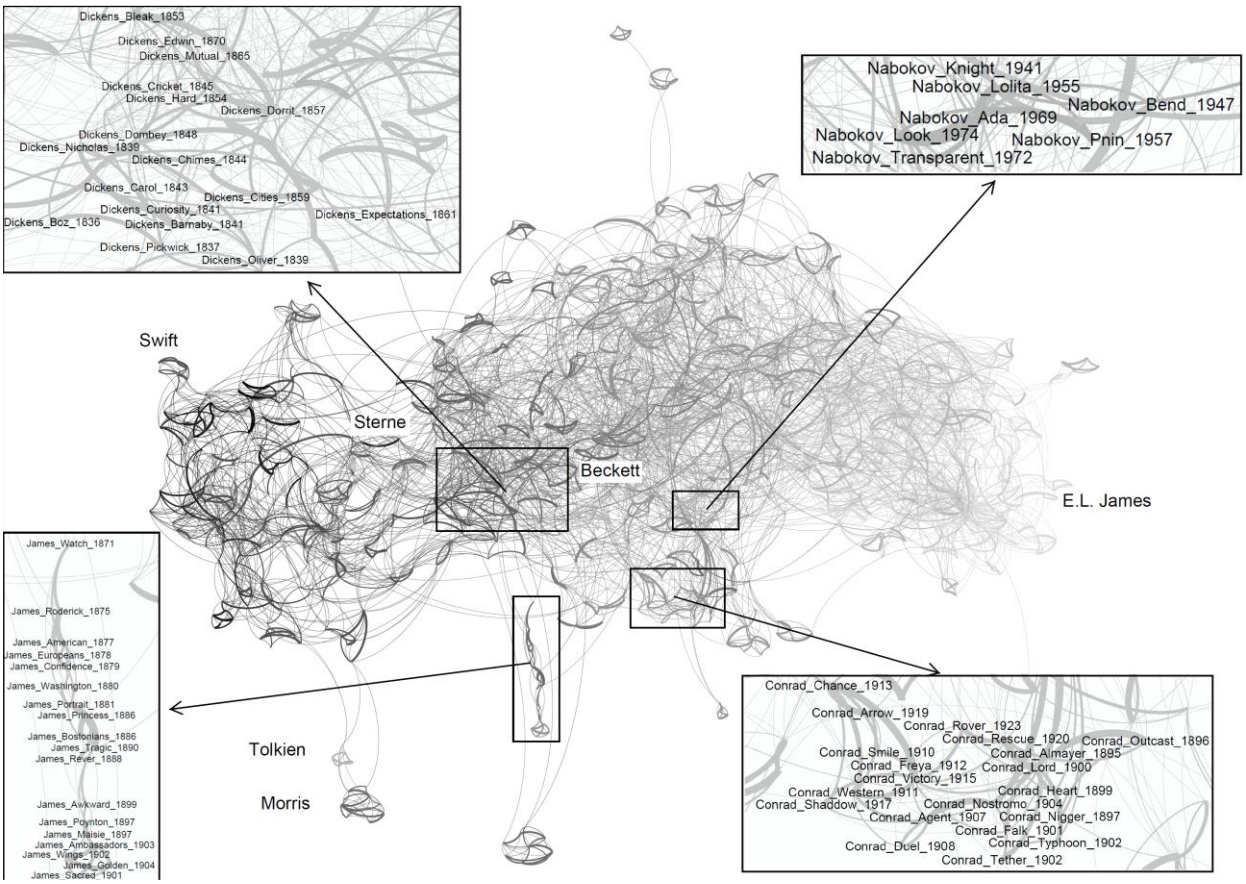

Figure 9. Network visualization of 1000 English novels between published between 1704 and 2013. Chronology is indicated by the transition from black (early) to light grey (late).

It would then seem that, in this context, the real interest of a network graph like the one in Fig. 9 is not where it agrees with literary history, but where it does not. The anomaly of counter-chronological Conrad has already been mentioned, but another Slavic ESL user, Nabokov, behaves in exactly the same way. Lawrence Sterne's *Sentimental Journey* and *Tristram Shandy* have been unseasonably removed to the right; this seems to illustrate quite well the very experimental nature of the latter work that culminated in Shklovsky calling its author "a radical revolutionary as far as *form* (emphasis added) is concerned" (Shklovsky 1990: 147). By contrast,

Beckett's trilogy leans left, and it has been called "archaic" by Harold Bloom (1988: 9). Nearby lies the magical land of Narnia, but that is even easier to explain as a throwback to archaized, or mediaevalized, fantasy. The other famous fantast, Tolkien, departs from the flow altogether as one of the outsiders; hanging on to the main body of the English novel network by the threads of his links to the historical romances of Bulwer-Lytton, he gravitates towards the poetic romances of William Morris – and that, too, seems to make a lot of sense.

5. Conclusions

The presentation of the above examples – the examples being necessarily simple and the presentation necessarily short – is intended to show how the tools and methods of Computational Stylistics can be used in the study of literature by any interested scholar who takes the trouble of learning the comparatively simple and comparatively user-friendly software. The authors have hoped to show that while Computational Stylistics sails into uncharted waters by becoming a partner in what Moretti has called distant reading (2015) and what Jockers has called macroanalysis (2013), there is much it can do to help in the traditional study of texts, whether it sticks to its earliest application in authorship attribution; whether it produces an image of literary history; or whether it helps to tell good student essays from bad.

It is interesting to see how the three levels of computational text analyses, the micro, the medium, and the macro, can produce results that create an added value at different levels of literary study. It is a very traditional task for scholars of literature to discern who wrote which part of a text; here, the computational stylist provides support – or correction – to what has been

established on the basis of such obvious sources as the correspondence and the reminiscences of Haggard and Lang. It sould be remembered that the importance of their *The World's Desire* in the body of popular fiction of their time can probably equal the recent media frenzy about the authorship of Harper Lee's *Go Set A Watchman* and *To Kill A Mockingbird*.

It is safe to assume, faced with the results of the study of *The Waves*, that Virginia Woolf did not count words she used to create the idiolects of her six "voices"; but the final outcome is as if she did with the sole purpose of diversifying her characters' voices. The most significant added value of this analysis to the existing body of academic criticism on Woolf is in fact nothing less that a heightened appreciation of the writer's genius: she has been able to produce character styles so distinctive that the differences between them are discernible through quantitative analysis; and that she has diversified them by age and by gender. The nightingale knows not how to read musical notes and how to measure intervals; but its song is brilliant nevertheless.

In a much more down-to-earth material, that of the high school essays, quantitative textual analysis produces results that are interesting for other reasons. Rather than providing an additional method of dealing with a large body of papers to be graded, it seems like a direct indication to students and teachers alike how to write essays; and this sort of investigation fits well into recent psycho- and sociolinguistic approaches exemplified by James Pennebaker's *The Secret Life of Pronouns* (2011).

One of the oldest tasks of literary studies, classification and systematics of authors, is well served by macroanalysis of the kind presented by the last example in this short survey. History of literature has been trying to put authors in groups and periods and tendencies without the

possibility to read them all; stylometric software reads books differently than humans, but at least it "reads" more of them than any single human – and it can graph the results of this "reading" according to much less impressionistic criteria than those criticized – and perhaps also those adopted – by such specialists on the function and the task of criticism as T.S Eliot or Terry Eagleton.

But it must be understood that computational stylistics has no quarrel with Eliots and Eagletons. It has been all three authors' experience that, when a computational stylist meets an open-minded traditional literary scholar, the twain come up with a new quality in textual analysis on any of the three scales discussed and exemplified in this paper. Computational Stylistics aims not at replacing Bloom with Burrows; it makes much more sense to bring them together for the benefit of our knowledge of literature, our potential in reading and understanding texts – and our appreciation of both the literary tradition and the individual literary talent.

References

Balossi, Giuseppina. 2014. *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves.* Amsterdam: Johns Benjamins.

Bloom, Harold. 1988. *Samuel Beckett's Molloy, Malone Dies, The Unnamable*. New York: Chelsea House.

Brainerd, Barron. 1980. "The Chronology of Shakespeare's Plays: A Statistical Study." *Computers and the Humanities* 14: 221-230.

Burrows, John F. 2010. “Never Say Always Again: Reflections on the Numbers Game.” In *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, ed. Willard McCarty, 13-35. Cambridge: Open Book Publishers.

Burrows, John F. 2006. “All the Way Through: Testing for Authorship in Different Frequency Strata.” *Literary and Linguistic Computing* 22 (1): 27–47.

Burrows, John F. 1994. “Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative.” *Research in Humanities Computing* 2: 1-33.

Burrows, John F. 1987. *Computation into Criticism*. Oxford: Clarendon Press.

Craig, Hugh. 1999. “Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything about Them?” *Literary and Linguistic Computing* 14(1): 103–113.

Craig, Hugh, and Arthur Kinney, eds. 2010. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press.

Eder, Maciej. 2015a. “Rolling Stylometry.” *Digital Scholarship in the Humanities* 30, advance access published 7 April 2015, doi: 10.1093/llc/fqv010.

Eder, Maciej. 2015b. “Visualization in Stylometry: Cluster Analysis Using Networks.” *Digital Scholarship in the Humanities* 30, advance access published 1 December 2015, doi: 10.1093/llc/fqv061.

Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch. 2015. “Revisiting Style, a Key Concept in Literary Studies,” *Journal of Literary Theory* 9(1): 25-52.

Hoover, David L. Forthcoming. “Argument, Evidence, and the Limits of Digital Literary Studies,” in Matthew Gold, ed., *Debates in the Digital Humanities*.

Hoover, David L. 2014. “Making Waves: Algorithmic Criticism Revisited,” DH2014, University of Lausanne and Ecole Polytechnique Fédérale de Lausanne, July 8-12.

Hoover, David L. 2013. “The Full-Spectrum Text-Analysis Spreadsheet,” *Digital Humanities 2013*, Lincoln, NE: Center for Digital Research in the Humanities, University of Nebraska, 226-29.

Hoover, David L. 2012. “The Tutor’s Story: A Case Study of Mixed Authorship,” *English Studies* 93(3): 324-339.

Hoover, David L. 2007. “Corpus Stylistics, Stylometry, and the Styles of Henry James” *Style* 41 (2): 174–203.

Jeffrey, Jill, David L. Hoover, and Mihye Han. 2013. “Lexical Variation in Highly and Poorly Rated U.S. Secondary Students’ Writing: Implications for the Common Core Writing Standards,” AERA 2013 Annual Meeting, San Francisco.

Jeffrey, Jill. 2010. “Voice, Genre, and Intentionality: an Integrated Methods Study of Voice Criteria Examined in the Context Of Large Scale-writing Assessment.” Diss. English Education. New York University.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana-Champaigne: University of Illinois Press.

Kestemont, Mike, Sara Moens, and Jeroen Deploige. 2015. “Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux.” *Digital Scholarship in the Humanities* 30(2): 199-224.

Moretti, Franco. 2015. *Distant Reading*. London: Verso.

Opas, Lisa Lena. 1996. “A Multi-Dimensional Analysis of Style in Samuel Beckett’s Prose Works.” *Research in Humanities Computing* 4: 81-114.

Pennebaker, James. 2011. *The Secret Life of Pronouns: What Our Words Say About Us*. New York, Bloomsbury Press.

Plasek, Aaron, and David L. Hoover. 2014. “Starting the Conversation: Literary Studies, Algorithmic Opacity, and Computer-Assisted Literary Insight,” DH2014, University of Lausanne and Ecole Polytechnique Fédérale de Lausanne, July 8-12.

Ramsay, Stephen. 2011. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

Rybicki, Jan, David L. Hoover, and Mike Kestemont. 2014. “Collaborative Authorship: Conrad, Ford, and Rolling Delta.” *Literary and Linguistic Computing* 29: 422–31.

Shklovsky, Victor. 1990. *Theory of Prose*. Elmwood Park: Dalkey Archive Press.

van Dalen-Oskam, Karina, and Joris van Zundert. 2007. “Delta for Middle Dutch—Author and Copyist Distinction in Walewein.” *Literary and Linguistic Computing* 22 (3): 345-62.